Multinomial logit bias reduction via Poisson log-linear model

Ioannis Kosmidis
Department of Statistical Science, University College
London, WC1E 6BT, UK
ioannis@stats.ucl.ac.uk

and

David Firth
Department of Statistics, University of Warwick
Coventry CV4 7AL, UK
d.firth@warwick.ac.uk

September 11, 2010

Abstract

It is shown how to obtain the bias-reducing penalized maximum likelihood estimator for the parameters of a multinomial logistic regression by using the equivalent Poisson log-linear model. This allows a simple and computationally efficient implementation of the reduced-bias estimator, using standard software for generalized linear models.
Key words: Jeffreys prior; leverage; logistic linear regression; Poisson trick.

1 Introduction

Use of the Jeffreys-prior penalty to remove the $O(n^{-1})$ asymptotic bias of the maximum likelihood estimator in exponential family models was developed in Firth (1993) and has been found to be particularly effective in binomial and multinomial logistic regressions (e.g., Heinze & Schemper, 2002; Bull et al., 2002, 2007). Implementation of the method in binomial and other univariate-response models is by means of a simple, iterative data-adjustment scheme (Firth, 1992). The purpose here is to extend such simplicity of implementation to multinomial models.

Suppose that observed k-vectors $y_1, \ldots, y_n$ of counts are realizations of independent multinomial random vectors $Y_1, \ldots, Y_n$. Let $m_r = \sum_{s=1}^{k} y_{rs}$ be the multinomial total and π_{rs} be the probability of the sth category for the multinomial vector Y_r, with $\sum_{s=1}^{k} \pi_{rs} = 1$ $(r = 1, \ldots, n; s = 1, \ldots, k)$. In multinomial logistic regression the log-odds of category s versus category k, say, for the rth multinomial vector is

$$\log\left(\frac{\pi_{rs}}{\pi_{rk}}\right) = \eta_{rs} = \beta_s^T x_r \quad (r = 1, \ldots, n; s = 1, \ldots, q)\,. \tag{1}$$

Here x_r is a corresponding vector of p covariate values, with first component unity if a constant is included in the model; and $\beta_s \in \Re^p$ is a vector of parameters for the sth category $(s = 1, \ldots, q)$, with $q = k-1$. Write $\gamma = (\beta_1^T, \ldots, \beta_q^T)^T$ for the row vector of all parameters. The linear predictor η_{rs} may then be expressed in the form $\eta_{rs} = \sum_{t=1}^{pq} \gamma_t g_{rst}$, where g_{rst} is the (s,t)th component of the $q \times pq$ matrix $G_r = I_q \otimes x_r^T$ with I_q the $q \times q$ identity matrix $(r = 1, \ldots, n; s = 1, \ldots, q)$.

Maximum likelihood estimation of γ may be performed either by maximizing the multinomial likelihood or by estimating via maximum likelihood the parameters $\theta = (\gamma^T, \phi_1, \ldots \phi_n)^T$ of a Poisson log-linear model with

$$\log \mu_{rs} = \zeta_{rs} = \phi_r + (1 - \delta_{sk})\beta_s^T x_r \quad (r = 1, \ldots, n; s = 1, \ldots, k)\,. \tag{2}$$

Here μ_{rs} $(r = 1, \ldots, n;\ s = 1, \ldots, k)$ represent the expectations of independent Poisson random variables Y_{rs}, and $\phi_1, \ldots, \phi_n$ are nuisance parameters; the Kronecker function δ_{sk} is equal to 1 when $s = k$ and zero otherwise.

The above equivalence was noted in Birch (1963), and Palmgren (1981) showed that the inverse of the Fisher information on $\beta_1, \ldots, \beta_q$ is the same in both representations under the restriction $\sum_{s=1}^k \mu_{rs} = m_r$ $(r = 1, \ldots, n)$ on the parameter space of the Poisson log-linear model. That restriction is automatically satisfied at the maximum likelihood estimate because if $l(\gamma)$ is the log-likelihood for the Poisson log-linear model, then $\partial l(\gamma)/\partial \phi_r = m_r - \tau_r$, where $\tau_r = \sum_{s=1}^k \mu_{rs}$ $(r = 1, \ldots, n)$.

2 Bias reduction via the log-linear model

In Firth (1992) it is shown that the bias-reducing adjusted score functions for the log-linear model (2) can be written in the form

$$U_t^* = \sum_{r=1}^{n} \sum_{s=1}^{k} \left(y_{rs} + \frac{1}{2} h_{rss} - \mu_{rs} \right) z_{rst} \quad (t = 1, \ldots, n + pq)\,. \tag{3}$$

Here z_{rst} is the (s,t)th component of the $k \times \{n + pq\}$ matrix

$$Z_r = \left[\begin{array}{c:c} G_r & 1_q \otimes e_r^T \\ \hdashline 0_{pq}^T & e_r^T \end{array}\right] \quad (r = 1, \ldots, n)\,,$$

with 0_{pq} being a pq-vector of zeros, 1_q a q-vector of ones, and e_r a vector of zeros with one in its rth component. The leverage quantity h_{rss} is the sth diagonal component of the $k \times k$ matrix $H_r = Z_r F^{-1} Z_r^T W_r$, where F is the Fisher information on θ and $W_r = \mathrm{diag}\,\{\mu_{r1}, \ldots, \mu_{rk}\}$ $(r = 1, \ldots, n)$. The matrix H_r is the $k \times k$, rth diagonal block of the asymmetric 'hat matrix' for model (2).

As in Firth (1992), expression (3) directly suggests an iterative procedure for solving the adjusted score equations: at the jth iteration, (i) calculate $h_{rss}^{(j)}$ $(r = 1, \ldots, n; s = 1, \ldots, k)$, where the superscript (j) denotes evaluation at the candidate estimate $\theta^{(j)}$ of the previous iteration and (ii) fit model (2) by maximum likelihood but using adjusted responses $y_{rs} + h_{rss}^{(j)}/2$ in place of y_{rs}, to get new estimates $\theta^{(j+1)}$. However, solution of the adjusted score equations $U_t^* = 0$ $(t = 1, \ldots, n + pq)$ does not result in the reduced-bias estimates of γ in the multinomial model. The reason is that by (3) the adjusted score equation for ϕ_r gives $\tau_r^* = m_r + \mathrm{tr}\{H_r^*\}/2$, with the star superscript denoting evaluation at the solution; this is in contrast to maximum likelihood, where the essential restriction $\hat{\tau}_r = m_r$ $(r = 1, \ldots, n)$ is automatic.

In order to construct a simple iterative procedure that does deliver the reduced-bias estimates of γ, it is convenient to re-express the model in terms of $\theta^\dagger = (\gamma^T, \tau^T)^T$, with $\tau^T = (\tau_1, \ldots, \tau_n)$, since then the necessary restriction is imposed directly by fixing the τ parameters at the corresponding multinomial totals. The log-linear model (2) is then re-written as a canonically-linked generalized nonlinear model,

$$\log \mu_{rs} = \log \tau_r + (1 - \delta_{sk})\eta_{rs} - \log \left\{ 1 + \sum_{u=1}^{q} \exp\left(\eta_{ru}\right) \right\} \quad (r = 1, \ldots, n; s = 1, \ldots, k)\,. \tag{4}$$

The variance and the third cumulant of Y_{rs} under the Poisson assumption are equal to μ_{rs} and the leverages h_{rss} are parameterization invariant. Hence, expression (13) in Kosmidis & Firth (2009) gives that the bias-reducing adjusted score equations using adjustments based on the expected information matrix take the form

$$U_t^\dagger = \sum_{r=1}^{n}\sum_{s=1}^{k}\left[y_{rs} + \frac{1}{2}h_{rss} + \frac{1}{2}\mu_{rs}\mathrm{tr}\left\{(F^\dagger)^{-1}\mathcal{D}^2\left(\zeta_{rs};\theta^\dagger\right)\right\} - \tau_r\pi_{rs}\right] z_{rst}^\dagger \quad (t = 1,\ldots,n+pq)\,,$$

where $F^\dagger$ is the expected information on $\theta^\dagger$, $\mathcal{D}^2\left(\zeta_{rs};\theta^\dagger\right)$ denotes the $(n+pq)\times(n+pq)$ Hessian matrix of ζ_{rs} with respect to $\theta^\dagger$, and $z_{rst}^\dagger$ is the (s,t)th component of the $k\times\{n+pq\}$ matrix

$$Z_r^\dagger = \left[\begin{array}{c:c} G_r - 1_q\otimes\left(\pi_r^T G_r\right) & 1_q\otimes\left(\tau_r^{-1}e_r^T\right) \\ \hdashline -\pi_r^T G_r & \tau_r^{-1}e_r^T \end{array}\right] \quad (r = 1,\ldots,n)\,,$$

with $\pi_r = (\pi_{r1},\ldots,\pi_{rq})^T$ and $\pi_{rs} = \mu_{rs}/\tau_r$ $(s = 1,\ldots,k)$.

After noting that $\mathcal{D}^2\left(\zeta_{rs};\theta^\dagger\right)$ does not depend on s and substituting for $z_{rst}^\dagger$ $(r = 1,\ldots,n; s = 1,\ldots,k)$, the adjusted score functions for γ take the simple form

$$U_t^\dagger = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} + \frac{1}{2}h_{rss} - \left(\tau_r + \frac{1}{2}\mathrm{tr}\{H_r\}\right)\pi_{rs}\right] g_{rst} \quad (t = 1,\ldots,pq)\,. \tag{5}$$

The following theorem provides some identities on the relationship between the matrix H_r and the $q\times q$, rth diagonal block of the asymmetric hat matrix for the multinomial logistic regression model (1). Denote the latter matrix by V_r.

Theorem 1 *Let v_{rsu} be the (s,u)th component of the matrix V_r $(r = 1,\ldots,n; s,u = 1,\ldots,k-1)$. If the parameter space is restricted by $\tau_1 = m_1,\ldots,\tau_n = m_n$ then*

$$h_{rss} = \pi_{rs} + v_{rss} - \sum_{u=1}^{q}\pi_{ru}v_{rus} \quad (s = 1,\ldots,q)\,,$$

$$h_{rkk} = \pi_{rk} + \sum_{s,u=1}^{q}\pi_{ru}v_{rus}\,,$$

where $\pi_{rs} = \mu_{rs}/\tau_r$ $(r = 1,\ldots,n; s = 1,\ldots,k)$.

Note that with the assumptions and identities in Theorem 1 it may immediately be seen that $\mathrm{tr}\{H_r\} = \mathrm{tr}\{V_r\} + 1$ $(r = 1,\ldots,n)$. Some algebra using the identities in Theorem 1 then yields that, under the restriction $\tau_r = m_r$ $(r = 1,\ldots,n)$, the adjusted score functions for γ in (5) take the form

$$U_t^\dagger = \sum_{r=1}^{n}\sum_{s=1}^{q}\left\{y_{rs} + \frac{1}{2}v_{rss} - \left(m_r + \frac{1}{2}\mathrm{tr}\{V_r\}\right)\pi_{rs} - \frac{1}{2}\sum_{u=1}^{q}\pi_{ru}v_{rus}\right\} g_{rst} \quad (t = 1,\ldots,pq)\,.$$

Using the results in Kosmidis & Firth (2009, p.797) on adjusted score functions for canonically-linked multivariate generalized linear models, further algebra shows that the above expression coincides with the adjusted score functions for the multinomial logistic regression model for any value of γ. The proof of Theorem 1 and details of the algebraic manipulations, which are straightforward but tedious, are in Kosmidis (2007, Appendices B.5 and B.6).

Hence the following iterative procedure applies: move from candidate estimates $\gamma^{(j)}$ to new values $\gamma^{(j+1)}$ by solving

$$0 = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} + \frac{1}{2}\tilde{h}_{rss}^{(j)} - \left(m_r + \frac{1}{2}\mathrm{tr}\left\{\tilde{H}_r^{(j)}\right\}\right)\pi_{rs}^{(j+1)}\right] g_{rst} \quad (t = 1, \ldots, pq)\,, \tag{6}$$

with $\tilde{h}_{rss}^{(j)}$ calculated under the restriction $\sum_{s=1}^{k}\mu_{rs}^{(j)} = m_r$. Directly from (5), the above iteration has a stationary point at the reduced-bias estimates of γ. Furthermore, from the form of the score functions for Poisson log-linear models, iteration (6) may be decomposed into the following steps:

1. set $\tilde{\phi}_r^{(j)} = \log\left\{\sum_{s=1}^{k}\mu_{rs}^{(j)} - \frac{1}{2}\mathrm{tr}\left\{H_r^{(j)}\right\}\right\} - \log\left\{1 + \sum_{s=1}^{q}\exp\left(\eta_{rs}^{(j)}\right)\right\} \quad (r = 1, \ldots, n)$,

2. use $\tilde{\theta}^{(j)} = (\gamma^{(j)}, \tilde{\phi}_1^{(j)}, \ldots, \tilde{\phi}_n^{(j)})$ to calculate a new value $\tilde{H}_r^{(j)}$ for the hat matrix $(r = 1, \ldots, n)$,

3. fit the log-linear model (2) by maximum likelihood but using the adjusted responses $y_{rs} + \tilde{h}_{rss}^{(j)}/2$ in place of y_{rs} to get new estimates $\theta^{(j+1)}$ $(r = 1, \ldots, n; s = 1, \ldots, k)$.

Note that H_r depends on the model parameters only through the Poisson expectations $\mu_{r1}, \ldots, \mu_{rk}$ $(r = 1, \ldots, n)$ and that the first step implies the rescaling of the current values of those expectations so that they sum to the corresponding multinomial totals. It is straightforward to implement this iteration using standard software for univariate-response generalized linear models; a documented program for the R statistical computing environment (R Development Core Team, 2010) is available from the second author upon request.

References

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B: Methodological* **25**, 220–233.

Bull, S. B., Lewinger, J. B. & Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Bull, S. B., Mak, C. & Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57–74.

Firth, D. (1992). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In *Computational Statistics I*, Y. Dodge & J. Whittaker, eds. Heidelberg: Physica-Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Kosmidis, I. (2007). *Bias Reduction in Exponential Family Nonlinear Models*. Ph.D. thesis, Department of Statistics, University of Warwick.

Kosmidis, I. & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793–804.

Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68**, 563–566.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.